2012

The oxytricha trifallax mitochondrial genome

Estienne C. Swart
Princeton University

Mariusz Nowacki
University of Bern

Justine Shum
Princeton University

Heather Stiles
Princeton University

Brian P. Higgins
Princeton University

See next page for additional authors

Genome Biology and Evolution

SMBE

The *Oxytricha trifallax* Mitochondrial Genome

Estienne C. Swart[1], Mariusz Nowacki[1,4], Justine Shum[1], Heather Stiles[1], Brian P. Higgins[1], Thomas G. Doak[2], Klaas Schotanus[1], Vincent J. Magrini[3], Patrick Minx[3], Elaine R. Mardis[3], and Laura F. Landweber[1,*]

[1]Department of Ecology and Evolutionary Biology, Princeton University

[2]Department of Biology, University of Indiana

[3]Genome Sequencing Center, Washington University School of Medicine

[4]Present address: Institute of Cell Biology, University of Bern, Bern, Switzerland

*Corresponding author: E-mail: lfl@princeton.edu.

Accepted: 8 December 2011

Data deposition: Mitochondrial plasmid—JN383842, Mitochondrial genome—JN383843, Macronuclear-encoded mt-DNA polymerase—JN383844, Macronuclear-encoded mt-RNA polymerase—JN383845.

Abstract

The *Oxytricha trifallax* mitochondrial genome contains the largest sequenced ciliate mitochondrial chromosome (~70 kb) plus a ~5-kb linear plasmid bearing mitochondrial telomeres. We identify two new ciliate split genes (*rps3* and *nad2*) as well as four new mitochondrial genes (ribosomal small subunit protein genes: *rps*- 2, 7, 8, 10), previously undetected in ciliates due to their extreme divergence. The increased size of the *Oxytricha* mitochondrial genome relative to other ciliates is primarily a consequence of terminal expansions, rather than the retention of ancestral mitochondrial genes. Successive segmental duplications, visible in one of the two *Oxytricha* mitochondrial subterminal regions, appear to have contributed to the genome expansion. Consistent with pseudogene formation and decay, the subtermini possess shorter, more loosely packed open reading frames than the remainder of the genome. The mitochondrial plasmid shares a 251-bp region with 82% identity to the mitochondrial chromosome, suggesting that it most likely integrated into the chromosome at least once. This region on the chromosome is also close to the end of the most terminal member of a series of duplications, hinting at a possible association between the plasmid and the duplications. The presence of mitochondrial telomeres on the mitochondrial plasmid suggests that such plasmids may be a vehicle for lateral transfer of telomeric sequences between mitochondrial genomes. We conjecture that the extreme divergence observed in ciliate mitochondrial genomes may be due, in part, to repeated invasions by relatively error-prone DNA polymerase-bearing mobile elements.

Key words: split genes, segmental duplication, genome expansion, linear mitochondrial plasmid, mobile elements, extreme mitochondrial divergences.

Introduction

Although ciliates are well known for their dimorphic macronuclear and micronuclear nuclear genomes, they also possess distinctive genomes in their mitochondria. The *Paramecium* and *Tetrahymena* mitochondrial genomes were among the first confirmed to be linear and to have their telomeric sequences identified (Suyama and Miura 1968; Goddard and Cummings 1975; Morin and Cech 1986, 1988). Ciliate mitochondrial genomes are both gene-rich and relatively large (20–60 kb) (Gray et al. 1998), though many mitochondrial genes remain unclassified (Pritchard et al. 1990; Burger et al. 2000; Brunk et al. 2003; Moradian et al. 2007), in part due to their extreme divergences from other eukaryotic mitochondrial genomes (Pritchard et al. 1990; Burger et al. 2000; Moradian et al. 2007). Split ribosomal RNA and *nad1* genes (Seilhamer, Gutell, et al. 1984; Seilhamer, Olsen, et al. 1984; Schnare et al. 1986, 1995; Heinonen et al. 1987; Pritchard et al. 1990; Burger et al. 2000) were also discovered in ciliate mitochondria. Some anaerobic ciliates contain hydrogen-producing organelles, or hydrogenosomes, that may derive from mitochondria, and the ciliate *Nyctotherus* (*Armophorea*) has a partially sequenced hydrogenosome genome (Akhmanova et al. 1998; Boxma et al. 2005). *Nyctotherus* may be more closely related to *Euplotes*

and *Oxytricha* (*Spirotrichea*) than to *Paramecium* and *Tetrahymena* (*Oligohymenophorea*), though this relationship still lacks convincing phylogenetic support (Ricard et al. 2008; de Graaf et al. 2009). Comparison of the mitochondrial and hydrogenosome genomes will permit examination of these relationships.

Sequencing and assembly of the macronuclear genome of *Oxytricha trifallax* also yielded most of its mitochondrial genome, which we completed by polymerase chain reaction (PCR) and sequencing. With the addition of this genome and the availability of complete mitochondrial genomes from two different ciliate phyla—the spirotrichs *Oxytricha* and *Euplotes* and oligohymenophorans *Paramecium* and *Tetrahymena*—detailed comparative genomic studies of ciliate mitochondria are now possible.

Materials and Methods

DNA isolation as described in Dawson and Herrick (1982) resulted in partially purified macronuclei, which were then used to produce libraries for Sanger and 454 sequencing from various populations of size-selected DNA. Mitochondrial DNA present in the libraries permitted recovery of the majority of the mitochondrial genome sequence; assembly with the Newbler (proprietary: www.454.com) produced two large mitochondrial contigs from pooled 454 and Sanger sequence data (currently represented by Contig4281.1 and Contig4553.1 from the 2.1.8 assembly). Additional sequences from PCR products amplified across the missing regions completed the mitochondrial genome sequence. We completed the mitochondrial genome assembly using these additional Sanger sequences, plus smaller contigs not originally merged in the two large contigs, using the Geneious software's assembler (Drummond et al. 2009).

To investigate the size of the mitochondrial plasmid, DNA was separated on an ethidium bromide-stained agarose gel, depurinated in-gel (0.25% HCl 15 min; washed in 0.4 M NaOH for 15 min) and transferred to Hybond XL membrane (Amersham) in 0.4 M NaOH using a Nytran TurboBlotter (Schleicher & Schuell). Labeled probe was generated by means of random priming (RadPrime, Invitrogen) of a wild-type *Oxytricha* strain JRB310 cloned PCR product. After overnight hybridization at 60 °C (0.5 M NaPO4, pH 7.2, 1% BSA, 1 mM EDTA, 7% SDS), the membrane was washed in 0.2× SSC with 0.1% SDS (30 min, 60 °C) and visualized on a GE Healthcare Storm 840 Phosphorimager.

To investigate the rRNA split gene structure in *O. trifallax*, RNA was isolated from *O. trifallax* strain JRB 310 using TRIzol according to the manufacturer's specifications (Invitrogen, Carlsbad, CA) and treated with Ambion DNA-free (Austin, TX) to remove contaminating DNA. Clean RNA was tailed with GTP in a standard reaction (1X NEB Buffer 2, 1 mM GTP, 5 μg RNA, and 2 U Poly (U) Polymerase [New England Biolabs; Ipswich, MA]) at 37 °C for 10 min. A reverse transcriptase reaction was performed using 1.6 μg tailed RNA and the Invitrogen SuperScript III First-Strand Synthesis System (Carlsbad, CA) with a UXR C12D primer (Horton and Landweber 2000). PCR was performed using 1X buffer, 10 mM each dNTP, 1.5 mM $MgCl_2$, 200 nM primer rnsaF (5′-TCGGAATGAACGCGAGCGGA-3′), 200 nM UXR anchor primer (5′-CATCATCATCATCTCGAGAATT-3′), ~80 ng cDNA, and 1.25 U Taq polymerase (Roche Applied Science; Indianapolis, IN). The PCR reaction conditions were: one cycle of 95 °C for 2 min, 35 cycles of 95 °C for 10 s, 58 °C for 10 s, and 72 °C for 30 s, before a final extension at 72 °C for 5 min. The same PCR was performed with an extension time of 60 s rather than 30 s for the primer pair rnsbF (5′-AGTTGCTCTGAAAGGTCGGACAA-3′) and UXR anchor, as well as the pair rnlaF (5′-CATTAAGTGGATGCCTATATATTGAATG-3′) and UXR anchor. Aliquots of each reaction were visualized in 2% agarose gels with SyBr Green using a Typhoon imager (GE Healthcare, Waukesha, WI). PCR products corresponding to expected sizes were cloned into plasmid pSC-Amp/Kan using the StrataClone PCR cloning kit (Stratagene; Santa Clara, CA) and sequenced.

Protein open reading frames (ORFs) were identified using a combination of Blast homology to either the NCBI nrdb or "mitochondrial" proteins from UniProt (UniProt Consortium 2011), when homology could be identified. ORFs were also predicted where no homology was detected by a custom python script, which provides a sliding window score for the probability of being a coding sequence and automatic ORF predictions in Geneious (Drummond et al. 2009). Since we do not know which start codons are employed in the *Oxytricha* mitochondrial genome, we have predicted start codons based on the start codons used in *Tetrahymena* (ATG, ATA, ATT, GTG, TTG). Predicted ORFs were at least 150-bp long.

We were able to detect homologs of most of the ciliate mitochondrial protein coding genes using conventional Blast-based homology searches. However, there are still a number of additional genes that are so divergent that they fall within or beyond the "twilight zone" of protein sequence similarity (Rost 1999), where Blast searches alone are unable to detect homology. An additional complication in these genomes is the presence of split genes, which may reduce sequence search sensitivity by shortening the regions available for local sequence alignment. We therefore used the more sensitive search technique provided by the HHpred web server (Soding et al. 2005), which uses a combination of PSI-Blast (Altschul et al. 1997) and HHsearch (an HMM-profile based search tool; Soding 2005). The latter tool is one of the fastest protein structure prediction tools with reasonable prediction accuracy (Hildebrand et al. 2009) and was recently useful in identifying an additional ciliate mitochondrial gene containing an *rps3* C-terminal domain (de Graaf et al. 2009). We also used Quickphyre (Kelley and Sternberg 2009) with default parameters to attempt to find homologs for a limited

number of ORFs. Two additional techniques assisted us in classifying "unknown" ORFs in *Oxytricha* and other ciliates: transitive homology relationships ("chains of homology"; Brunk et al. 2003) and the inference of orthology based on extensive synteny within spirotrichs and oligohymenophorans (and, to a lesser degree, between these classes).

We used tRNAscan-SE (Lowe and Eddy 1997) with default parameters and the "mitochondrial/chloroplast" source option to identify tRNAs in the *Oxytricha* mitochondrial genome.

Quikfold, from the UNAfold package on the DINAMelt (Markham and Zuker 2005) server, was used to predict the mitochondrial plasmid DNA hairpins with the temperature set to 20 °C, $[Na^+]$ 1 M, and $[Mg^{2+}] = 0$ M.

Transmembrane helices were predicted using THMM2 (Krogh et al. 2001) with default parameters.

PAML version 3.15 was used to estimate d_n/d_s ratios (Yang 1997), in pairwise run mode with standard parameters, except that the genetic code was set to translation table = 4.

Results

Structure of the Ciliate Mitochondrial Chromosome

Approximately 70 kb *O. trifallax* mitochondrial genome shares a number of structural features with the existing ciliate mitochondrial genomes (fig. 1; GenBank accession JN383842). As in the *Tetrahymena* and *Euplotes* mitochondrial genomes, the *Oxytricha* mitochondrial genes are predominantly or exclusively arranged in two transcriptional directions, diverging from an approximately central location, whereas both the *Paramecium* mitochondrial genome and *Nyctotherus* hydrogenosome genome have one primary direction of transcription (fig. 1). The *Oxytricha* mitochondrial DNA has a relatively high AT content (76%, excluding telomeres) as is typical for mitochondria in general (Gray et al. 2004). To date, there seems to be little taxonomic consistency in mitochondrial genome base composition within ciliates: *Tetrahymena pyriformis* has an AT content similar to *O. trifallax* at 79% (Burger et al. 2000), whereas *P. tetraurelia* (Pritchard et al. 1990) and *E. minuta* (de Graaf et al. 2009) have a considerably lower AT content at 59% and 64%, respectively. The *Nycotherus* hydrogenosome DNA (GenBank accession: GU057832.1) is also less AT rich (58.5%).

In all ciliate mitochondrial genomes, including that of *Oxytricha*, there is either a central (in *Tetrahymena* [Burger et al. 2000], *Euplotes* [de Graaf et al. 2009], and the hydrogenosome of *Nycotherus* [de Graaf et al. 2011]) or terminal (in *Paramecium* [Pritchard et al. 1990]) region bearing low sequence complexity repeats. In *Paramecium* (Goddard and Cummings 1975, 1977; Pritchard and Cummings 1981) and *Tetrahymena* (Arnberg et al. 1974), the AT-rich region coincides with the origin of DNA replication (in *Tetrahymena* it is contained within the largest mitochondrial ORF, *ymf77*, encoding translated [Smith et al. 2007] 1,386 aa protein of unknown function. The *Tetrahymena paravorax* mitochondrial genome contains the longest AT-rich stretch, ~1 kb of 96.5% AT sequence adjacent to the major site of change in transcription direction (Moradian et al. 2007). The central repeats in *T. pyriformis*, *T. pigmentosa*, and *T. malaccensis* are shorter, at a few hundred base pairs each. In *Oxytricha*, the central region is a ~140-bp long stretch of pure AT, composed of degenerate repeats of the unit (written as a POSIX regular expression): ((AAAT) + (AT)+){4,}) which contains stretches of potentially self-complementary repeats, typically palindromes such as TATA, TATATA, and TATATATA. The presence of DNA structures that would be refractory to DNA polymerase is indicated by our difficulty in amplifying across this region using conventional PCR. The *Euplotes* > 1 kb central repeat region is more GC-rich than that of the other ciliate mitochondrial regions (~83.5% AT) and is comprised of semipalindromic 18-bp repeats (TANNATGTATACATNNTA). *Paramecium* possesses pure AT repeats in its terminal region (TATTTATTAAAATAAAAAAATATAAATATATTAA). *Nyctotherus*'s hydrogenosome repeat is considerably more GC rich (46.7% AT) than all the other ciliate mitochondrial genome repeats. Since the hydrogenosome genome is still incomplete, it is possible that terminal AT-rich repeats are missing from this genome.

The *O. trifallax* mitochondrial genome is capped by telomeres consisting of 35-bp repeats of CGACTCCTCTATCCTCATCCTAGACTCCGCTTACT, with an unknown repeat number (the longest assembled mitochondrial telomeric repeat consists of approximately 15 repeat units) and appears to be linear, like the mitochondrial genome of *Tetrahymena*. As in *Tetrahymena* and *Paramecium*, we have found no macronuclear genome–encoded telomerase RNA with this repeat. The telomeric repeat units are in the same size range as those for a variety of *Tetrahymena* species (Morin and Cech 1988) (31–53 bp) but more GC-rich (51.4% vs. 26.0–40.0%). No sequence data for similar telomeric repeats has been published for *Paramecium*, for which a different end-replication model, based on cross-links between the two DNA strands, has been proposed (Pritchard and Cummings 1981; Nosek et al. 1998) nor for *Euplotes* or *Nyctotherus*.

Like *Tetrahymena*, the *Oxytricha* mitochondrial genome also has a terminal inverted repeat (TIR) just inside the telomeric repeats, comprised of a somewhat smaller region (~1,800 bp; 87.8% identical, including a 96 bp indel) (fig. 1) than *Tetrahymena's* (~2,680 bp). This region is roughly bounded by a *trnC* and a putative *trnC* pseudogene (*trnC-ψ*). The *Tetrahymena* inverted repeat is largely comprised of the large subunit ribosomal RNAs and also contains tRNAs, including *trnL* paralogs, whereas *Oxytricha's* appears to be largely comprised of protein-coding ORFs of unknown function. The presence of potentially unrelated terminal inverted duplicated genes in both *Oxytricha* and *Tetrahymena* suggests that this region may be an important source for gene duplications in these genomes. Aside from ciliates, TIRs

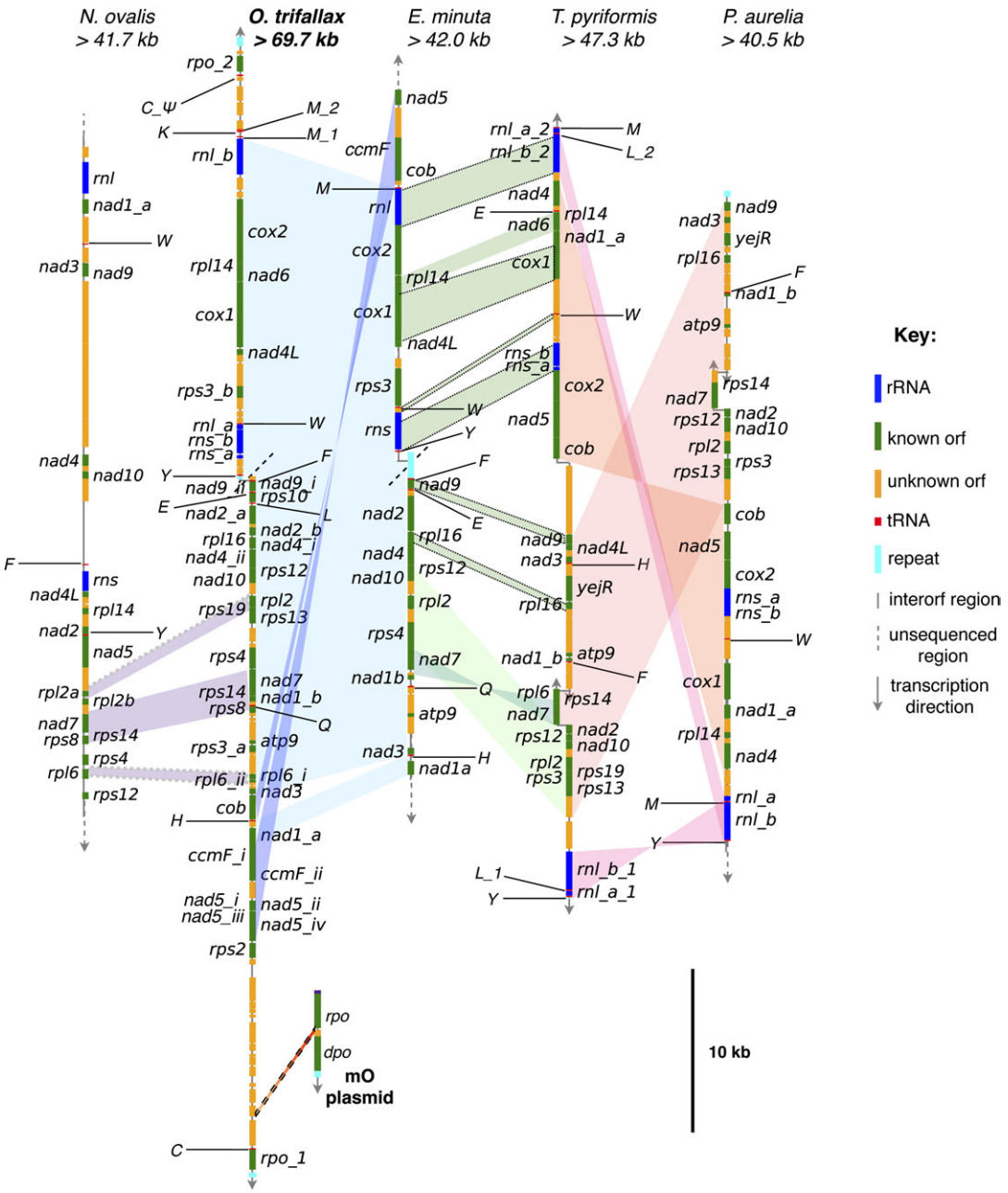

Fig. 1.—Gene map of the *Oxytricha trifallax* mitochondrial genome (GenBank accession: JN383842) in comparison to that of representative ciliate mitochondrial genomes (*Euplotes minuta*—GQ903130; *Tetrahymena pyriformis*—AF160864; *Paramecium tetraurelia*—NC_001324) and the *Nyctotherus ovalis* hydrogenosome genome (GU057832.1). Unknown ORFs are as currently annotated in GenBank, without genes that we have newly classified. Split genes are suffixed with an underscore followed by an alphabetic character. *Oxytricha trifallax* ORFs with a lowercase roman numeral are fragments we predict to belong to the same gene but which are possibly artificially split by sequencing errors. The undetermined lengths of the central repeat regions of *O. trifallax* and *E. minuta* are indicated by a dashed diagonal line. tRNAs are indicated by single letter amino acid codes. Collinearity between the genomes is indicated by pale-colored regions, including the collinear, but interrupted, single genes that are demarcated by finely dashed lines. The ~250-bp region shared by the mitochondrial mO plasmid (JN383843) and the primary genome of *O. trifallax* is indicated by a red and black dashed line.

are characteristic of many linear mitochondrial genomes from diverse eukaryotes, including yeasts, such as *Pichia pijperi* (~1.8 kb) and *Williopsis saturnus* (~1.9 kb) (Dinouel et al. 1993); chytridomycete fungi, such as *Hyaloraphidium curvatum* (~1.4 kb) (Forget et al. 2002); cnidarians, such as *Hydra oligactis* (Kayal and Lavrov 2008); slime molds, such as *Physarum polycephalum* (~0.6 kb) (Takano et al. 1994); and unicellular green algae, such as *Chlamydomonas*

reinhardtii (580 bp) and *Polytolmella parva*, which have a well-conserved TIR of ~1.5 kb shared by the four ends of its bipartite mitochondrial genome (Fan and Lee 2002). TIRs appear to be a common characteristic not only of mitochondrial genomes but of many linear eukaryotic and bacterial plasmids as well (Meinhardt et al. 1990, 1997) and have been proposed to be a solution to the end-replication problem for linear mitochondrial molecules (Dinouel et al. 1993; Vahrenholz et al. 1993).

Ciliate Mitochondrial Genome Synteny

The levels of interclade synteny and intraclade synteny (fig. 1) of the mitochondrial genomes of the four genera representing two ciliate classes, *Spirotrichea* (*Oxytricha* and *Euplotes*) and *Oligohymenophorea* (*Tetrahymena* and *Paramecium*), are consistent with current taxonomic classification. There is extensive collinearity within both the spirotrich and oligohymenophorean mitochondrial genomes, with the amount of collinearity between the mitochondrial genomes of *Tetrahymena* and *Paramecium* (Burger et al. 2000) comparable to that between *Oxytricha* and *Euplotes*.

The mitochondrial genomes of *Tetrahymena* and *Paramecium* are largely collinear, with the exception of one large inversion and translocation (*nad9-ymf76* in *Tetrahymena*; *nad9-orf105* in *Paramecium*) (Burger et al. 2000); whereas the *Oxytricha* and *Euplotes* genomes are largely collinear in the core region, from *nad3* to *rnl*. The decrease in collinearity between classes reflects the more ancient divergence of the two classes. The ciliate mitochondrial gene order is fairly static, considering the ancient evolutionary divergences of these species (as much as 2 billion years since the divergence of oligohymenophorans from spirotrichs [Wright and Lynn 1997]). We propose that the relatively static mitochondrial genome synteny could be exploited as an additional useful classification tool at higher ciliate taxonomic levels.

A large region of six mostly adjacent genes, *rps4/ymf76* (*T. pyriformis* or *ymf81* and *ymf85* in *P. tetraurelia*), *rps13*, *rps19*, *rpl2*, *nad10*, and *rps12*, is present in all sequenced ciliate mitochondrial genomes (fig. 1). In *T. pyriformis* and *P. tetraurelia*, this region also includes "*rps3*" (Brunk et al. 2003) (now classified as *rps3_a*; see "protein-coding genes") and an unknown gene. The *nad7* and *rps14* genes are also adjacent in all the ciliate mitochondrial genomes but are in the inverted transcription direction in the oligohymenophoran mitochondrial genomes. After accounting for the inversion in the *Paramecium* mitochondrial genome, extensive collinearity is still present between the oligohymenophoran and spirotrich genomes (*trnM*, *rnl*, *rpl14*, *nad6*, *cox1*, *trnW*, *rns*, the AT-rich replication origin/transcription initiation region, *nad9*, and *rpl16*, and the largely adjacent *rps4*, *rps13*, *rps19*, *rpl2*, *nad10*, *rps12* genes). Taken as a whole, extensive collinearity between the *Oxytricha*, *Euplotes* and *Tetrahymena* mitochondrial genomes—plus the observation of the most extensive genome reduction in *Paramecium*—points to *Paramecium* possessing a derived ciliate mitochondrial genome form.

There is limited collinearity (e.g., *rpl2*, *nad7*, *rps14*, *rps8*, *rpl6*) between the spirotrich mitochondrial genomes and the *Nyctotherus* hydrogenosome genome that may reflect the tumult of the change from a mitochondrion to a hydrogenosome.

Mitochondrial Genome Gene Content

Protein-Coding Genes

As can be seen in table 1, with the exception of a few gene losses, ciliate mitochondrial genomes share largely the same complement of known protein-coding genes. The partially sequenced hydrogenosome genome from *Nycotherus* contains a subset of *Oxytricha* mitochondrial protein-coding genes: *Nyctotherus* has lost genes required for aerobic metabolism, in particular the cox genes (Boxma et al. 2005; de Graaf et al. 2011). Details of the identification and annotation of previously undiscovered or unannotated proteins in ciliate mitochondrial genomes and the *Nycotherus* hydrogenosomal genome are provided in supplementary table 1 (Supplementary Material online). In total, we have been able to annotate seven previously unidentified genes in *Euplotes*, six in *Tetrahymena* and *Paramecium*, and three in *Nyctotherus*.

Oxytricha's complement of small ribosomal proteins, in particular, is fairly complete, compared with other protist repertoires (Gray et al. 2004): all ribosomal proteins except for *rps1* and *rps11* have been identified in all four sequenced ciliate mitochondrial genomes. We found homologs for all but one (*ymf61*) of the *Tetrahymena* putative ribosomal proteins, for which no homologs were found using conventional Blast searches (Brunk et al. 2003). The fact that three of the newly classified ribosomal proteins (*rps4*, *7*, *10*) are commonly encoded in protist mitochondrial genomes (Gray et al. 2004) but were missing from the *Tetrahymena* mitochondrial proteome survey (Smith et al. 2007) (which would have detected nuclear versions of these proteins, had they been transferred to the nucleus), instills confidence in these gene predictions. With the addition of these small subunit ribosomal proteins, most of the common mitochondrially encoded protist ribosomal proteins (Gray et al. 2004) appear to have been discovered, in the mitochondrial or nuclear genomes in ciliates.

Our annotations of a number of previously unannotated *Tetrahymena* mitochondrial-encoded genes, plus the availability of proteomic and bioinformatic identification of nuclear-encoded *Tetrahymena* mitochondrial genes (Smith et al. 2007), indicates that the remainder of the unknown *Tetrahymena* ORFs are largely nonribosomal, in agreement with a previous study (Brunk et al. 2003). These unknown ORFs could be novel mitochondrial proteins or proteins that

Table 1
Ciliate Mitochondrially Encoded Genes

	Oxytricha trifallax	*Euplotes minuta*	*Nyctotherus ovalis*	*Tetrahymena pyriformis*	*Paramecium tetraurelia*
nad1 a	*	*	* (no split determined)	*	*
nad1 b	*	*		*	*
nad2 a	*	*	*	*ymf65*	*ymf65_a + b*
nad2 b	*	*	*	*nad2*	*nad2*
nad3	*	*	*	*	*
nad4	*	*	*	*	*
nad4L	*	*	*	*	*
nad5	*	*	*	*	*
nad6	*	*	*ord236*	*	*
nad7	*	*	*	*	*
nad9	*	*	*	*	*
nad10	*	*		*	*
cob	*	*		*	*
cox1	*	*		*	*
cox2	*	*		*	*
atp9	*	*		*	*
ccmF/yejR	*	*		*	*
rps2	*		*orf262*		
rps3 a	*	*		*ymf64*	*ymf64*
rps3 b	*	*orf190*		*rps3*	*rps3*
rps4	*	*		*ymf76*	*ymf81 + 85*
rps7	*	*orf170*		*ymf63*	*ymf63*
rps8	*	*orf125*	*	*ymf74*	*ymf84*
rps10	*	*orf111*		*ymf59*	*ymf59*
rps12	*	*	*	*	*
rps13	*	*orf102*		*	*
rps14	*	*orf49+*	*	*	*
rps19	*	*orf155*	*orf199*	*	*
rpl2	*	*	*	*	*
rpl6	*	*	*	*	
rpl14	*	*	*	*	*
rpl16	*	*		*	*

NOTE.—MCiliate mitochondrially encoded genes determined in this study and in previous studies (Pritchard et al. 1990; Burger et al. 2000; Brunk et al. 2003; Moradian et al. 2007; de Graaf et al. 2009, 2011; Barth and Berendonk 2011). *rps2*, *rps7*, *rps8*, and *rps10* are newly discovered ciliate mitochondrial proteins, with previously unrecognized orthologs in other species.

have diverged beyond our ability to detect homology to known proteins. The sequencing of additional ciliate mitochondrial genomes (particularly ciliate classes other than spirotrichs and oligohymenophoreans) may be beneficial in identifying the remaining unidentified genes in ciliate mitochondrial genomes, especially if HMM-HMM profile comparison tools (such as HHpred) are used to improve the information content and quality of the alignments underlying the query HMM profiles.

Split Protein-Coding Genes in Ciliates

rps3. The *Euplotes rps3* is unusually long (767 and 768 amino acids for *E. minuta* and *Euplotes crassus*, respectively) in comparison to the *rps3* orthologues found in the *Oxytricha* (~349 aa), *Tetrahymena* (330 aa), and *Paramecium* (234 aa) and was show to contain the C-terminal domain of *rps3* in the 5′-terminal half of this gene (de Graaf et al. 2009) (fig. 2). The 3′ half of the *Euplotes* gene has no detectable similarity to *rps3*. In *Oxytricha*, this same gene is divided into a shorter, 5′-terminal portion containing the C-terminal *rps3* domain, followed by a longer portion of unknown function. We identified an *Oxytricha* homolog to a gene previously classified as *rps3* (Burger et al. 2000) in *Tetrahymena* and *Paramecium* but disputed as such (Brunk et al. 2003). As for *Tetrahymena* and *Paramecium* and unlike *Euplotes* (de Graaf et al. 2009), HHpred predicts with high probability (5.1×10^{-06} for *Oxytricha*) that an N-terminal *rps3* domain is present in the mitochondrial genome, in an ORF that we label *rps3_a*. It is possible that the *rps3* N-terminal domain is encoded in a missing portion of the *Euplotes* mitochondrial genome. It therefore appears that this is another split gene present in most, if not all, sequenced ciliate mitochondrial genomes. Accordingly, the previously disputed *rps3* (N-terminal) can now be called *rps3_a*, and the recently classified C-terminal *rps3* portion, *rps3_b*, consistent with the split gene nomenclature in Burger et al. (2000).

The long gene annotated as *rps3* in *Euplotes* may represent either a novel gene fusion or an incorrect annotation

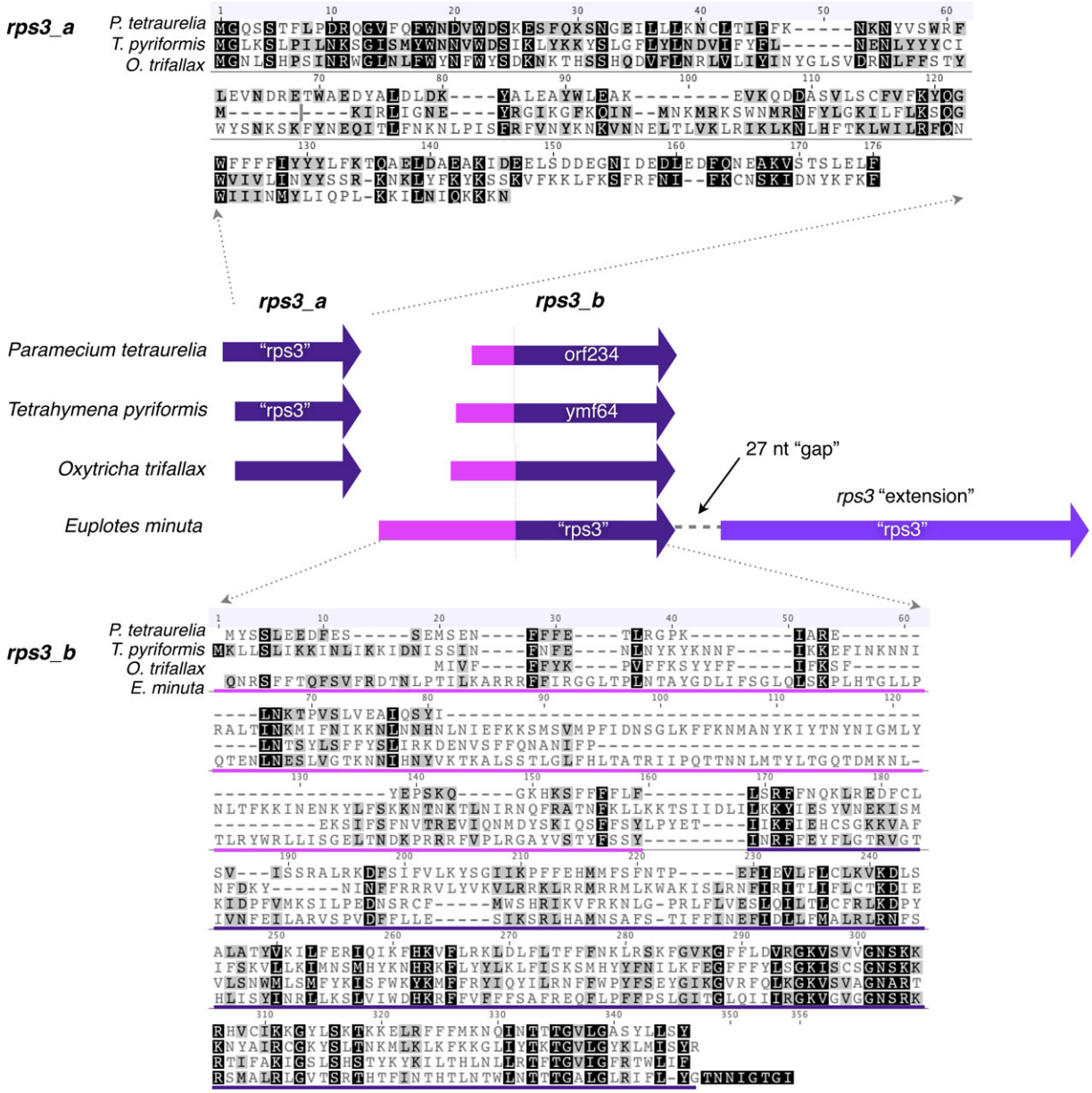

Fig. 2.—*rps3* genes in ciliate mitochondrial genomes (*Euplotes minuta*—GQ903130; *Tetrahymena pyriformis*—AF160864; *Paramecium aurelia*—NC_001324). The *rps3_a* and *rps3_b* multiple sequence alignments are indicated above and below a schematic representation of the split *rps3* genes. Regions with substantial sequence similarity are indicated in dark purple, whereas those that are poorly conserved are indicated in pink; the *rps3_a* and *rps3_b* parts are on distant loci. The *rps3* extension annotated as a part of this gene in *Euplotes* does not align to any of the other *rps3* sequences. Multiple sequence alignments were generated with Muscle (Edgar 2004) with default parameters.

due to sequencing errors. The sum of the lengths of the *Oxytricha*, *Tetrahymena*, and *Paramecium rps3* domains is roughly consistent with typical UniProt *rps3* entries (e.g., ~480 aa for *Tetrahymena*), although there is considerable length variation in *rps3* among species—even within fungi alone (Sethuraman et al. 2009). Some of the *rps3* genes we inspected appear to be missing a domain (e.g., we found no N-terminal *rps3* domain in *Schizosaccharomyces pombe*). It appears that there is some flexibility in the intervening *rps3* domain spacer: in humans the intervening spacer between the N- and C-terminal domains contains a single-stranded nucleic acid binding domain (KH domain) required for stable NF-κB regulatory complex binding, an extra-ribosomal function (Wan et al. 2007). In plants, the N- and C-terminal *rps3* domains are separated by a domain of unknown function (Smits et al. 2007). One other case of a split arrangement of the N- and C-terminal *rps3* domains has been documented in the slime mold *Dictyostelium*, where the domains

are separated by long peptide sequences of unknown function (Iwamoto et al. 1998; Smits et al. 2007). Unlike *Dictyostelium*, the ciliate split *rps3* ORFs are located some distance from one another, with multiple genes separating them, and in both *Oxytricha* and *Euplotes*, they are encoded on opposite strands.

nad2. Part of this gene was previously identified in *Paramecium*, *Tetrahymena*, and *Euplotes*, but its length varies greatly: the *Tetrahymena* and *Paramecium* "*nad2*" genes are unusually short, just 166 and 178 aa long, respectively. By contrast, the shortest curated *nad2* gene in UniProt is 346 aa in *Branchiostoma*, whereas the longest is 538 aa in *Ustilago*. *Euplotes crassus* (391 aa) and *E. minuta* (722 aa) are thought to possess N-terminal extensions with no substantial sequence similarity to other *nad2* genes (de Graaf et al. 2009). HHpred searches using the original *Tetrahymena nad2* ORF revealed that this ORF contains only a C-terminal portion of *nad2*.

We found an ORF (372 aa) in the *Oxytricha* mitochondrial genome that appears to be weakly similar to a "putative nonribosomal" *Tetrahymena* protein, *ymf65* (BlastP *e* value 0.028 to *T. malaccensis nad2*) and to the *Hydra* and *Phytophthora nad2* genes (*e* values of 1.2 and 4.8 to NCBI's nrdb, respectively). In *Oxytricha*, this region is separated from an ORF (167 aa) with BlastP hits to the existing annotated ciliate *nad2* entries in GenBank (*e* values of 10^{-06} to 10^{-05}), by a 60 aa unknown ORF. HHpred predictions for either the 372 aa *Oxytricha* ORF or *Tetrahymena ymf65* correspond to the N-terminal half of *nad2*. *ymf65* was previously predicted to have ten transmembrane helices (Brunk et al. 2003), the same number we obtained for the 372 aa *Oxytricha* ORF using THMM2 (Krogh et al. 2001) (fig. 3).

In *Paramecium*, the *nad2* N-terminal region appears to be further split into two ORFs: *ymf65_a* and *ymf65_b*. Like the transmembrane helix sums, the length sums of the ORFs corresponding to *nad2* for *Tetrahymena*, *Paramecium*, and *Oxytricha* (526, 518, and ~543 aa) are within the range of eukaryotic *nad2* lengths in UniProt protein sequences.

The annotated *nad2* gene from the *Nyctotherus* hydrogenosome genome (GenBank accession: AJ871267.1) encodes the C-terminal *nad2* portion and is a similar length (166 aa) to its *Oxytricha*, *Tetrahymena*, and *Paramecium* orthologs. An ORF preceding the *Nyctotherus nad2* C-terminal region does not share substantial sequence similarity with—and is approximately 100 amino acids shorter than—the *Oxytricha*, *Tetrahymena*, and *Paramecium nad2* N-terminal regions. The N-terminal half of the *Nyctotherus nad2* appears to be the ORF currently annotated as orf371.

The sums of the transmembrane helices from the N- and C-terminal *nad2* ciliate gene portions, for all but *Euplotes* (14 for *Tetrahymena* and *Oxytricha*, 13 for *Paramecium*, and 15 for *Nyctotherus*), are in accord with the number predicted for most nonmetazoan eukaryotes (13–14, e.g., *Arabidopsis thaliana* [UniProt: O05000], *Dictyostelium discoideum* [UniProt: O21048]. The overall spacing of the helices in the N-to-C terminal concatenated sequences of *nad2* from *Oxytricha*, *Nyctotherus*, and *Tetrahymena* and *Paramecium* THMM2 profiles are also in good agreement.

In *E. minuta*, the entire ORF annotated as *nad2* is predicted by THMM2 (Krogh et al. 2001) to contain 17 transmembrane helices, whereas *E. crassus* appears to have 12. The annotated *nad2* from *E. crassus* is just half the length (391 aa) of *E. minuta* (774 aa) and appears to encode only the C-terminal portion of the *E. minuta nad2* (57.8% pairwise identity). The pairwise alignments of the concatenated translations of the ORFs upstream of the shorter *E. crassus nad2* (*orf175* and *orf147*) to the remaining N-terminal portion of *E. minuta nad2* are only 18.7% identical, suggesting that these ORFs are either highly divergent or unrelated. Judging from the lengths and transmembrane numbers of *nad2* from *Oxytricha*, *Tetrahymena*, and *Paramecium*, the *E. crassus nad2* is also a split gene, though we have not identified with certainty the location of the missing part. Barring sequencing and annotation errors, these *Euplotes nad2* genes may indicate that *nad2* can be split or fused in multiple ways. It therefore appears that *nad2* may be split to different extents in different ciliate species (possibly independent evolutionary events), though a common *nad2* N/C-terminal split appears to be shared by *Tetrahymena*, *Paramecium*, *Oxytricha*, and *Nyctotherus*.

A *nad2* split gene is also present in angiosperm mitochondrial genomes (Malek et al. 1997). In these plants, *nad2* is joined by the transsplicing of a group II intron (Binder et al. 1992). We have not detected any group II introns in the ciliate mitochondrial genomes by scans of the RFAM (Griffiths-Jones et al. 2005) group II intron model with Infernal (Nawrocki et al. 2009). In *Oxytricha*, we think it is unlikely that RNA editing removes all the stops necessary to join the *nad2* ORFs (at least two stop codons would need to be eliminated or read through). Instead, it appears that, like *nad1* (Seilhamer, Gutell, et al. 1984; Seilhamer, Olsen, et al. 1984; Heinonen et al. 1987; Pritchard et al. 1990; Schnare et al. 1995; Burger et al. 2000), this gene is not transspliced (supported by cDNA PCR results, data not shown), and therefore, the gene pieces are translated as separate subunits that require co-assembly to form the functional protein structure.

Split rRNA Genes

In both *Tetrahymena* and *Paramecium*, the large and small subunit rRNA genes are further split (Seilhamer, Gutell, et al. 1984; Seilhamer, Olsen, et al. 1984; Heinonen et al. 1987; Pritchard et al. 1990; Schnare et al. 1995; Burger et al. 2000) into large and small portions. The current *T. pyriformis* and *P. tetraurelia* GenBank annotations present two different structures for the *rns* split: *Tetrahymena* has a short *rns_a* followed by a long *rns_b*, whereas *Paramecium* has a long

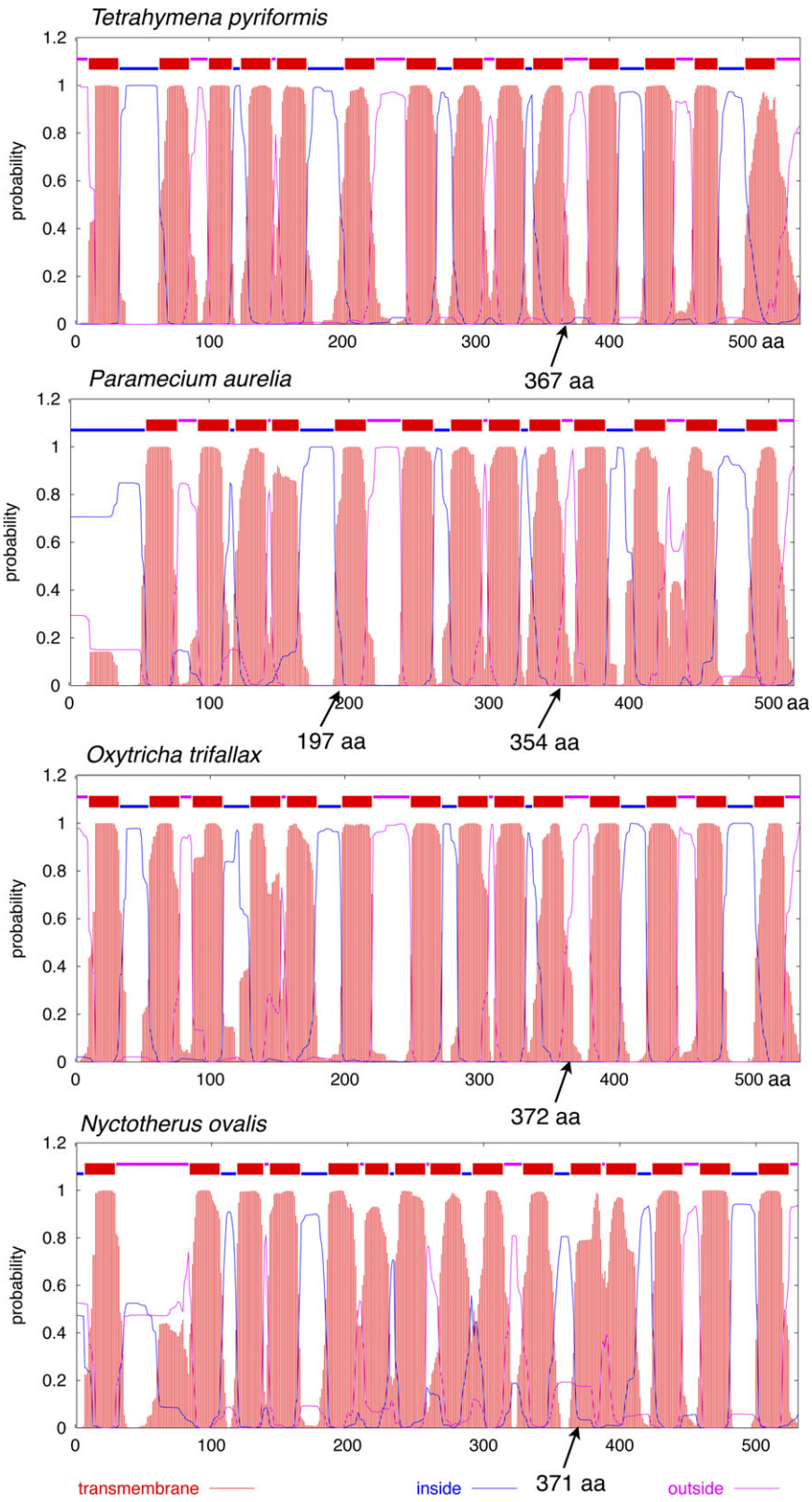

Fig. 3.—THMM2 transmembrane profile predictions for the concatenated *nad2* split ORFs from *Tetrahymena pyriformis* (AF160864), *Paramecium tetraurelia* (NC_001324), *Oxytricha trifallax* (JN383842), and *Nyctotherus ovalis* (GU057832.1). THMM2 posterior probabilities are given on the *y* axis; the *x* axis length is in amino acids. Concatenation points are indicated by arrows.

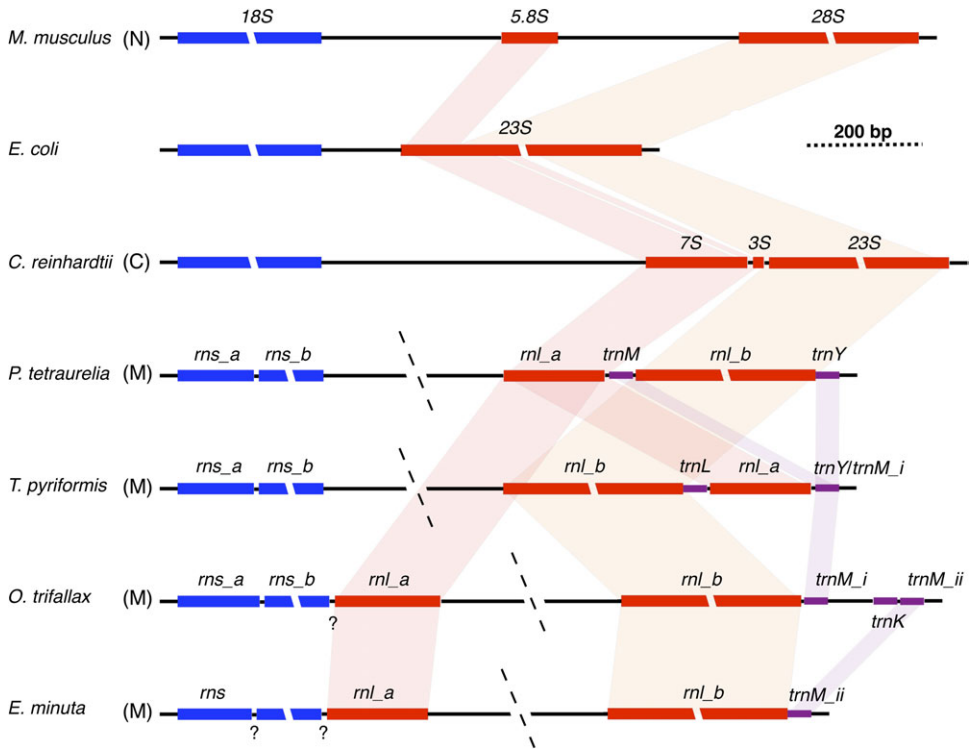

Fig. 4.—Comparison of ciliate split rRNA genes. Solid red and blue bars represent small and large subunit rRNA coding sequences, respectively, drawn approximately to scale; discontinuous red and blue bars represent longer sequences that have been compressed due to figure space constraints; black lines represent intervening sequences which are approximately halved relative to the coding sequences, except the central, large discontinuous intervening region in ciliates, indicated by the dashed line, which represents an extensive, primarily protein-coding, region; tRNA genes are purple. The duplicated large subunit region of *Tetrahymena pyriformis* is represented here by a single locus: the primary difference between the two loci is a different tRNA succeeding *rnl_a* (*trnY* and *trnM_i*). Homology between the different segments is indicated by pastel-colored parallelograms. Question marks indicate the lack of experimental evidence in *Oxytricha* or *Euplotes* supporting or rejecting gene splits. GenBank accession numbers for the rRNA loci are: *Mus musculus* nuclear genome rRNA (BK000964.1); *Escherichia coli* genome (NC_000913); *Chlamydomonas reinhardtii* chloroplast genome (NC_005353.1); *Paramecium tetraurelia* (NC_001324), *T. pyriformis* (AF160864.1), *Euplotes minuta* (GQ903130), and *Oxytricha trifallax* (JN383842) mitochondrial genomes.

rns_b followed by a short *rns_a*. However, the experimental results in *Paramecium* were misinterpreted and can be interpreted instead as *rns* having the same structure in both ciliates (Schnare et al. 1995). For *Euplotes*, no split rRNAs were detected using local sequence alignments of portions of the identified rRNA sequence (de Graaf et al. 2009), but no experimental support was provided for this conclusion. Northern analysis in *Nyctotherus* suggests that its SSU rRNA may be fragmented into three pieces of 1.7 kb, 750 bp, and 600 bp (Akhmanova et al. 1998).

Our inspection of alignments to the *Tetrahymena* and *Paramecium* split rRNA genes, as well as unpublished expressed sequence tag (EST) data, indicate that *Oxytricha* has the same splits identified in the LSU and SSU genes for *Paramecium* and *Tetrahymena* (fig. 4). EST data suggest that a long AT-rich DNA spacer (~141 bp; 96% AT) divides *Oxytricha rns* into two subunits, as in *Tetrahymena* and *Paramecium*, because no EST reads span this region. We also verified this split by obtaining RACE products corresponding to the 3′ end of *rns_a* (supplementary fig. 1, Supplementary Material online). Sequence alignments suggest that the ribosomal RNA fragment following *rns_b* is orthologous to the *rnl_a* fragment of *Tetrahymena*'s SSU RNA, which is physically separated from the remaining *Oxytricha rnl_b* fragment. A short AT-rich tract (~32 bp; 91.6% AT) between the *rns_b* and *rnl_a* in *Oxytricha* is also poorly covered by expression data (Swart EC, Landwebe LF, unpublished data) and hence a likely splitting point. Alignments of *rns* from the different ciliate species, including the *Nyctotherus rns* gene (see NCBI AJ871267.1) suggest that an *rnl_b* portion is located at the end of the gene annotated as *rns* in *Euplotes*. Furthermore, there are discrepancies in length of the annotated *Euplotes rnl* (2,230 bp) in comparison to that of other ciliate *rnl*s (~2,550–2,600 bp) (see table 3), These lines of evidence suggest that the *Euplotes* ribosomal LSU is split like that of *Tetrahymena*, *Paramecium*, and *Oxytricha*. Alignments of the ciliate SSU regions suggest that *Euplotes* could also possess the split SSU. At least one split (between *rnl_a* and *rnl_b*) is shared by all the ciliates and likely arose in their common ancestor. Additional

Table 2
Ciliate Mitochondrially Encoded tRNAs

	Oxytricha trifallax	*Euplotes minuta*	*Nyctotherus ovalis*	*Tetrahymena pyriformis*	*Paramecium tetraurelia*
trnC	*				
trnE	*	*		*	
trnF	*	*	*	*	*
trnH	*	*		*	
trnK	*				
trnL	*			*	
trnM i	*			*	*
trnM ii	*	*			
trnQ	*	*			
trnW	*	*	*	*	*
trnY	*	*	*	*	*

experimental evidence is necessary to determine whether the SSU split is present in *Euplotes* and hence common to all known ciliate mitochondrial genomes.

Mitochondrial Genetic Code and Transfer RNA Genes

In contrast to the different nuclear genetic codes in *Euplotes* versus *Oxytricha*, *Paramecium*, and *Tetrahymena*, these species all seem to share the same mitochondrial genetic code: a small variation on the "mold, protozoan mitochondrial code" ("table 4" according to NCBI Blast tables), with a single stop codon (UAA) and a single unused codon (UAG), whereas UGA encodes tryptophan (Pritchard et al. 1990; Burger et al. 2000; de Graaf et al. 2009). Our Blast searches for release factors in *Oxytricha*, *Paramecium*, and *Tetrahymena* identified a single nuclear-encoded mitochondrial peptide chain release factor (mtRF). As is typical for mitochondrial release factors, this release factor is more closely related to bacterial peptide chain release factor 1, which is involved in recognition of the UAA and UAG codons (Scolnick et al. 1968), than to standard eRFs. In bacteria and some eukaryotic mitochondria, RF2 recognizes the UAA and UGA codons. Codon reassignment from UGA as a stop codon to tryptophan is common in many eukaryotes and has occurred independently in numerous lineages (Massey and Garey 2007). Since both RF1 and RF2 recognize UAA, there is functional redundancy at this codon, which means that RF2 can be lost if UGA is no longer used as a stop codon. This reassignment is more likely to occur in small genomes, such as those in mitochondria, in a loss-and-regain-of-codon model (Massey and Garey 2007).

The *O. trifallax* mitochondrial genome encodes 11 tRNA genes corresponding to 10 different tRNA species (see table 2). Like the protein-coding genes of known function, the *O. trifallax* mitochondrial tRNA collection is a superset of those previously discovered in ciliate mitochondrial genomes to date. *Tetrahymena pyriformis* has eight tRNA genes corresponding to seven tRNA species (Burger et al. 2000); *E. minuta* has at least seven tRNAs corresponding to seven species; *P. tetraurelia* has four tRNA genes corresponding to four species (Pritchard et al. 1990; Burger et al. 2000). *Nycotherus* only has three hydrogenosome-encoded tRNAs, *trnF*, *trnW*, and *trnY*, the same tRNAs that are common to all ciliate mitochondrial genomes. *Oxytricha trifallax* possesses two versions of the *trnC* gene (73.5% identical, including gaps) located at both nontelomeric ends of the terminal repeat: a short form (73 bp) that tRNAscan-SE recognizes as *trnC* and a longer form (80 bp) that tRNAscan-SE designates an unknown tRNA. RNAfold (Hofacker 2003) using default parameters predicts that the former forms the characteristic cloverleaf secondary structure, whereas the latter may form a noncloverleaf structure and hence could be a tRNA pseudogene.

tRNAscan-SE (Lowe and Eddy 1997) also predicts two *trnM* genes for *O. trifallax*. One appears to be orthologous to the *trnM* in *Tetrahymena* and *Paramecium* (*trnM_i*), whereas the other (*trnM_ii*) more closely resembles the "*trnM*" of *Euplotes* (60% pairwise sequence identity vs. 55% for *Euplotes trnM* to *Oxytricha trnM_i*). The *Tetrahymena*/*Oxytricha*/*Paramecium trnM_i* orthologs share a characteristic, truncated D stem and loop (Heinonen et al. 1987; Schnare et al. 1995), whereas the *Oxytricha*/*Euplotes trnM_ii* appears to be structurally more similar to typical initiator and elongator *trnM*s from other eukaryotic mitochondrial genomes, such as those from *Reclinomonas*. The *trnM_ii* is the least well-conserved tRNA of the orthologs shared between *Euplotes* and *Oxytricha*, at 56.2% pairwise similarity versus an average identify of 70.3% (minimum

Table 3
Ciliate Split Ribosomal RNA Segment Lengths (bp)

	Oxytricha trifallax	*Euplotes minuta*	*Tetrahymena pyriformis*	*Paramecium tetraurelia*
rnl a	301	281[a]	289	280
rnl b	2,289	2,230	2,279	2,315
rnl	2,590	2,511[a] (2,230)	2,568	2,595
rns a	200	206[a]	208	212[b] (1,477)
rns b	1,426	>1,431[a]	1,407	1,415[b] (204)
rns	1,626	1,637[a] (2,257)	1,615	1,627

[a] Length estimates from our sequence alignments.
[b] Estimates based on an experimental reassessment (Schnare et al. 1986) of the original *P. tetraurelia* rRNA split gene structure (Seilhamer, Olsen et al. 1984). Bracketed values indicate the lengths of the rRNA segments as annotated in the mitochondrial genomes deposited in GenBank (*E. minuta*—GQ903130 and *P. tetraurelia*—NC_001324).

Table 4
Signal Peptide Prediction for Putative Macronuclear-Encoded Mitochondrial Polymerases

	Macronuclear-Encoded mtDNA Polymerase		Macronuclear-Encoded mtRNA Polymerase	
	Mitoprot	Predotar	Mitoprot	Predotar
Oxytricha trifallax	0.99	0.84	0.97	0.86
Tetrahymena thermophila	0.96	0.46	0.24	0.64
Paramecium tetraurelia	0.41	0.76	0.98	0.78

NOTE.—Mitochondrial signal prediction probabilities are shown. Annotated macronuclear-encoded mitochondrial DNA and RNA polymerase chromosomes have been deposited in GenBank (JN383844 and JN383845). The relatively low probabilities for *T. thermophila* and *P. tetraurelia* gene predictions may be due to incorrect gene start predictions (for instance, the *T. thermophila* mtRNA polymerase has a 200 aa extension compared with both the *P. tetraurelia* and *O. trifallax* gene predictions).

66.7%) for the remaining pairs of tRNA orthologs (*trnE*, *trnF*, *trnH*, *trnQ*, *trnW, trnY*; locARNA [Will et al. 2007]), suggesting that this gene has either been evolving under relatively relaxed constraints or positive selection since divergence from the common ancestor of these two ciliates. The divergence between the two *Oxytricha trnM*'s (55.9%) is greater than that between typical eukaryotic initiator and elongator *trnM*s; for instance, in *Reclinomonas*, these genes are 67.1% identical, suggesting that there may be substantial functional divergence between the two classes of ciliate mitochondrial *trnM*s. Given the substantial divergence of ciliate *trnM*s, we are hesitant at this point to ascribe the role of initiator or elongator tRNA to any of the ciliate *trnM*s.

A Linear Mitochondrial Plasmid

During inspection of the *Oxytricha* mitochondrial genome assembly, we discovered an additional large contig (~4.9 kb) possessing an internal region with substantial sequence similarity—251 bp at 82% identity—to one of the *Oxytricha* mitochondrial contigs (position indicated on fig. 1). We subsequently confirmed that this contig is a ~5.28-kb linear plasmid (fig. 5*a*), which we call mO (GenBank accession: JN383842). The 251-bp region appears to be a "footprint" of a past recombination event between the plasmid and mitochondrial genome. The region of similarity is reminiscent of a 473-bp sequence shared by the *Physarum* mitochondrial genome and its linear plasmid mF, which has been shown to permit integration of the linear plasmid into the mitochondrial genome via homologous recombination (Takano et al. 1992).

Consistent with a mitochondrial origin, the plasmid genes appear to use the same genetic code as the ~70 kb *Oxytricha* mitochondrial chromosome. Translation with either the standard or ciliate genetic codes would produce much shorter ORFs, due to in-frame UGA codons. However, we do note that the tryptophan codon bias (the telltale signature of the mitochondrial genetic code) of these ORFs (12 UGA vs. 9 UGG) is much weaker than that of known or "ciliate-specific" mitochondrial ORFs (i.e., those that appear to have orthologs in other ciliates; 229 UGA vs. 20 UGG). The deviation from the standard *Oxytricha* mitochondrial tryptophan codon usage suggests that either this plasmid may be a relatively recent acquisition that has not yet acquired the standard mitochondrial codon usage or that selection on codon usage is weaker on this plasmid.

The plasmid contains two large ORFs. The 3′ ORF encodes a linear mitochondrial plasmid/phage-like DNA polymerase that is easily identifiable by Blast (BlastP best-hit *e* value 8×10^{-15} to a *Fusarium proliferatum* linear plasmid DNA polymerase [GenBank accession: YP_001718360]). The DNA polymerase has the characteristic "DTDS" residues of the DNA Pol B family active site (Hopfner et al. 1999) and appears to be a member of phage and linear plasmid (including mitochondrial plasmids) DNA polymerases (Kempken et al. 1992) (see next section).

The 5′ ORF has no convincing Blast or HHpred hits but appears to be an RNA polymerase based on a QuickPhyre (Kelley and Sternberg 2009) prediction: the best QuickPhyre prediction (*e* value 0.41; estimated precision 80%) is to an X-ray crystal structure of a phage N4 virion RNA polymerase, related to the T7-like RNA polymerases (Kazmierczak et al. 2002). Linear plasmids, including mitochondrial ones, bearing both an RNA and DNA polymerase are a typical form (Meinhardt et al. 1990; Handa 2008). We also discovered that the longest ORF in the TIR of the *Oxytricha* mitochondrial genome, encoding a 411 aa protein, is predicted by Phyre (*e* value 0.67; 80% precision) to be related to the same phage N4 RNA polymerase as the mO RNA polymerase (65% precision, *e* value 1.4). The sequence similarity shared between the mO RNA polymerase and the TIR protein is a meager 17% (global pairwise alignment, gap opening, and extensions penalties of 12 and 3, using the BLOSUM62 matrix).

We were unable to find any evidence of protein homologs of the mitochondrial plasmid ORFs in either *Paramecium* or *Tetrahymena* mitochondrial genome data. Linear mitochondrial plasmids have been identified in *P. caudatum*, *P. jenningsi* and *P. micromultinucleatum* (Endoh et al. 1994; Tsukii et al. 1994) but none of their sequences has been published. In a low coverage *Stylonychia lemnae* genome assembly, we were able to find a telomere-lacking contig with a substantial TBlastN hit to the DNA polymerase ORF (contig12708 [http://lamella.princeton.edu/blast/getseq.cgi?454AllContigs.fna&contig12708]; *e* value 7×10^{-18}). The entire contig appears to be translated with the same mitochondrial genetic code as the one used in *Oxytricha* and not the *Oxytricha*/*Stylonychia* macronuclear genetic code or standard genetic code (which would introduce two premature stop codons). This *Stylonychia* Blast hit suggests that mitochondrial plasmids may be present in other spirotrichous ciliates.

The mO 5′ terminal ~250 bp contains three types of short semipalindromic repeats (fig. 5*b*), which are capable

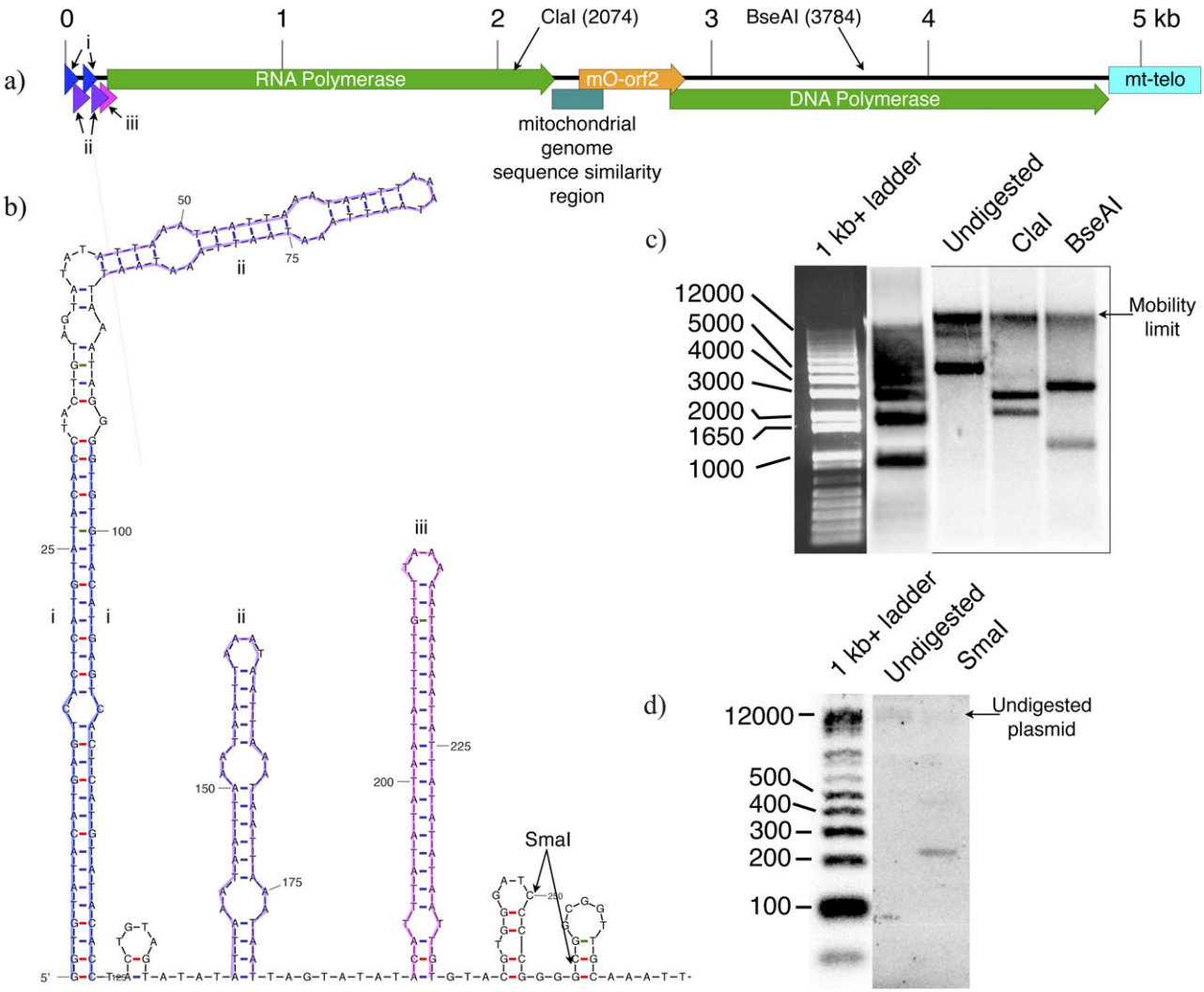

Fig. 5.—The *Oxytricha trifallax* linear mitochondrial plasmid. The linear plasmid is approximately drawn to scale. The three ORFs are indicated by arrows; the putative integration site is indicated in teal. The three classes of hairpin-forming sequences are indicated by triangles. Quikfold (Markham and Zuker 2005) structures predicted for the 5′ end of the plasmid. Southern analysis of the linear plasmid; the digestion product lengths are in agreement with a linear form of the mitochondrial plasmid; the probe also hybridizes to the mitochondrial genome in the mobility limit as it includes the region of sequence similarity shared by the plasmid and the main mitochondrial genome. Southern analysis to infer the length of the 5′ end of the plasmid.

of producing stem-loop structures of similar size to those of the *Physarum* linear plasmid mF 205-bp TIR region. The 3′ end of mO is capped by the same type of telomeric repeats as the main mitochondrial genome. At least one example of short (5 bp), telomere-like repeats on a linear plasmid has been reported for the fungus *Fusarium oxysporum* (Walther and Kennell 1999) (the *F. oxysporum* genome does not contain telomeres since it is circular [Marriott et al. 1984]). The *P. polycephalum* mitochondrial plasmid (mF) has longer—144 bp—subterminal repeats following the plasmid TIRs and is capable of in vivo linearization of the circular mitochondrial genome by recombining with it (Sakurai et al. 2000). Unlike *Physarum*, the *Oxytricha* mitochondrial genome's telomeric repeats appear to be established, rather than new extensions from the plasmid. This suggests that we have found the first possible case of a stable transfer of a telomere between a mitochondrial genome and a plasmid.

Putative Macronucleus-Encoded Mitochondrial DNA and RNA Polymerases

While no DNA or RNA polymerase genes have been documented in the mitochondrial genomes of *Tetrahymena* and *Paramecium*, there are nuclear-encoded candidates for these genes, which appear to have orthologs in the *Oxytricha* macronuclear genome as well. We sought to clarify the relationship between these putative mitochondrial polymerases and the plasmid-encoded polymerases found in *Oxytricha*.

Mitochondrial DNA polymerases are largely unknown or uncharacterized in most eukaryotes (Shutt and Gray 2006) with the exception of humans and yeast (Kaguni 2004). The opisthokont (metazoa/fungi) mitochondrial DNA polymerase (Pol gamma) is a Pol A family DNA polymerase, like bacterial DNA Pol I but a distinct (Lecrenier and Foury 2000) and divergent member of this family (20–25% identity relative to the *Escherichia coli* Klenow fragment [Lecrenier et al. 1997]). However, Pol gamma does not appear to exist in many eukaryotic lineages (Burgers et al. 2001) and so a different mitochondrial DNA polymerase must take its place in these organisms. In *Arabidopsis* and the red alga *Cyanidioschyzon merolae*, a single putative mitochondrial DNA polymerase, which is not orthologous to Pol gamma, is targeted to both mitochondria and plastids (Elo et al. 2003; Moriyama et al. 2008). These polymerases are more similar to bacterial DNA Pol I polymerases than to Pol gamma (Mori et al. 2005) and form a distinct clade—"plant organellar polymerases" (POPs; Moriyama et al. 2008)—comprised of diverse eukaryotic members, including the amoebozoan, *D. discoideum* (Shutt and Gray 2006) the heterokont *Phytophthora ramorum*; diatoms; plants; red alga; and the ciliates *T. thermophila* (GenBank accession: XP_001014571) and *P. tetraurelia* (GenBank accession: XP_001431083) (Moriyama et al. 2008). The ciliate POPs, including that of *Oxytricha* (GenBank accession: JN383844), appear to possess characteristic mitochondrial targeting signal peptides (table 4) and, therefore, are putative ciliate mitochondrial DNA polymerases.

A T-odd phage RNA polymerase homolog was identified for *T. pyriformis* during searches for homologues to the yeast mitochondrial and T3/T7 RNA polymerases (Cermakian et al. 1996). A complete homolog of the *T. pyriformis* sequence was predicted in the *T. thermophila* macronuclear genome assembly (Eisen et al. 2006; GenBank accession: XP_001013489) and subsequently discovered in the *T. thermophila* mitochondrial proteome (Smith et al. 2007). Both *P. tetraurelia* (GenBank accession: XP_001435950) and *O. trifallax* (GenBank accession: JN383845) homologs also exist in the respective macronuclear genome assemblies. Mitochondrial target signal prediction software (Mitoprot [Claros and Vincens 1996] and Predotar [Small et al. 2004]) predicts that the *Tetrahymena*, *Paramecium*, and *Oxytricha* proteins are mitochondrially targeted (table 4).

The *Oxytricha* linear plasmid RNA and DNA polymerases are so extremely divergent in comparison to the macronucleus -encoded putative mitochondrial RNA and DNA polymerases that conventional multiple sequence alignments of these genes are unreliable. Mitochondrial polymerases in the nuclear genomes of other organisms appear to be more closely related to the putative mitochondrial, nuclear-encoded ciliate polymerases and not to the plasmid polymerases, which appear to be a secondary acquisition.

An Abundance of Subterminal Unknown Open Reading Frames

Both ends of the *Oxytricha* mitochondrial genome—corresponding to ~5 kb and ~14.5 kb (or in total just over $^1/_4$ of the *Oxytricha* mitochondrial genome length)—contain almost exclusively ORFs without obvious homologues in any known organism or in the *Oxytricha* macronuclear genome. These regions constitute over half (54%) of the total unknown ORF length in the *Oxytricha* mitochondrial genome. The sum of these end regions accounts for the majority of the size difference between the smaller *Tetrahymena* and *Euplotes* mitochondrial genomes and *Oxytricha*'s larger one. The structure of the *Oxytricha* mitochondrial genome resembles the core *Euplotes* mitochondrial genome, with the exception of one large translocated region—the *cob*-to-*nad5* gene block—located at opposite ends of these mitochondrial genomes (fig. 1), with two large blocks of unknown ORFs appended to either end.

Segmental duplications are evident from a dot plot of the ~14.5 kb telomeric end (fig. 6). The largest of these duplications is closest to the telomeric end and is ~1,450 bp long with ~91% pairwise identity. An ~170-bp region represents the sequence that has been duplicated most often. Pairwise identities relative to the first repeat (from the telomeric end) from this region decrease with increasing distance: 93.4%, 88.9%, and 74.7%. If we assume that these regions are evolving approximately neutrally, then the duplications closest to the telomeric end are younger than the distal ones. This suggests that the ~14.5-kb region arose, in part, through successive expansions resulting in up to three successive terminal duplication events. Curiously, the 251-bp segment shared by the plasmid and mitochondrial genome is located near (~120 bp from) the end of the most recent duplication of these repeats.

The largest duplication within the 14.5-kb end contains at least one long ORF (~600 bp), which appears to be evolving under similar levels of evolutionary constraint ($d_n/d_s = 0.288$) to that of the pair of ORFs from the TIRs ($d_n/d_s = 0.302$). In *Tetrahymena*, both nonterminal duplications of *nad9* and the TIRs appear to be maintained by concerted evolution (Brunk et al. 2003), which is also likely to be the case for the *Oxytricha* TIR regions. These levels of constraint are somewhat lower than those we recently reported for the micronuclear-encoded *Oxytricha* TBE transposase paralogs (Nowacki et al. 2009). The synonymous substitution rate is also lower in these genes—0.095 for ~600 bp duplicated ORF and 0.181 for the TIR ORF pair—than the TBE transposase paralogs (0.287 average), indicating that these genes have either duplicated more recently than the TBEs (since ciliate mitochondrial genes evolve more rapidly than their nuclear genes) and/or that substitution has been suppressed by concerted evolution. Assuming no translocations and that the strength of concerted evolution either declines with

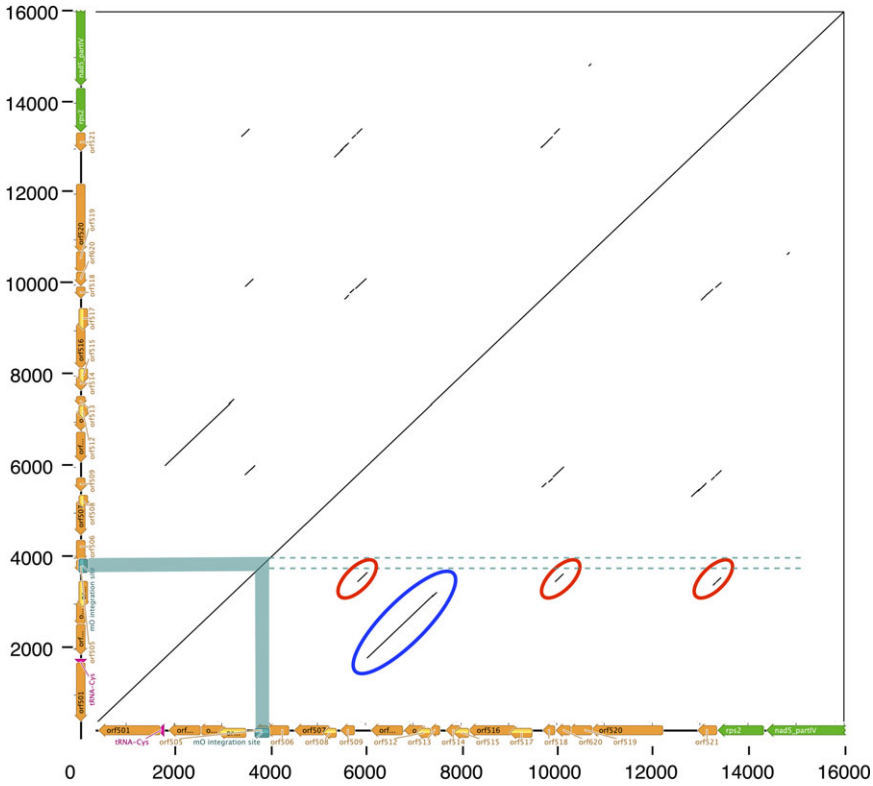

Fig. 6.—Segmental duplications in the 16-kb subterminal mitochondrial region of the *Oxytricha* mitochondrial genome long arm. The dotplot was generated by Dotmatcher (Emboss) (Rice et al. 2000) with a window size of 50 and threshold of 150. The axis scales are in base pairs. Along the axes, unknown ORFs are colored orange and known protein-coding genes are green. The ~170-bp quadruplication is enclosed by red ellipsoids; the 1,450-bp duplication is enclosed by a blue ellipsoid. The footprint of the mO plasmid is indicated in teal; the close proximity of this site to the end of the first region that is quadruplicated is indicated by dashed lines.

distance from the telomeres or remains approximately the same throughout the genome, the lower overall substitution rates in the 600 bp duplicated ORF pair, relative to the TIR ORF pair, suggest that these duplications arose both internally to, and after, the TIR ORF pair.

The pattern of sequence conservation in the mitochondrial terminal regions suggests both that purifying selection acting upon a duplicated ORF permitted the detection of duplications and that selective constraints have been lost in many of the surrounding ORFs leading to pseudogene formation. Two lines of evidence suggest pseudogenization: 1) the intervening regions between ORFs are longer in the terminal unknown ORF regions (~179 bp mean; ~178 bp standard deviation [SD]; excluding zero length regions) than the central region (~53 bp mean; 63 bp SD); 2) inter-ORF regions constitute ~19.5% of the terminal unknown ORF regions, whereas these spacers constitute ~5% of the *Oxytricha* mitochondrial genome excluding the terminal unknown ORF regions, a figure close to that of typical tightly packed ciliate mitochondrial genomes, such as those of *Tetrahymena* and *Euplotes* (Burger et al. 2000; de Graaf et al. 2009); 3) ORFs in the unknown ORF regions are shorter (419 bp mean; 310 bp SD) compared with those from the central region (501 bp mean; 412 bp SD).

In the 14.5-kb subterminal region, we also noticed an overabundance of tryptophan UGG (15) versus UGA (43) anticodons, in comparison to 20 UGG versus 229 UGA anticodons in conserved or ciliate-specific ORFs in the central region (the ~5-kb subtelomeric region has 0 UGG and 17 UGA tryptophan anticodons). This deviation from the standard tryptophan codon usage in the larger subterminal region could be an indication either of a relatively recent incorporation of foreign genetic material derived from the mO plasmid and/or relaxed constraint associated with possible pseudogene formation.

Discussion

The *Oxytricha* mitochondrial genome, at ~70 kb, is the largest ciliate mitochondrial genome sequenced to date. It is approximately 22–30 kb larger than the other completely sequenced ciliate mitochondrial genomes; mitochondrial genomes of the distantly related oligohymenophorans *P. tetraurelia* and *T. pyriformis* are ~40 kb (Pritchard et al. 1990)

and ~47 kb (Burger et al. 2000), respectively; the more closely related spirotrichous ciliate *E. crassus* is <48 kb (de Graaf et al. 2009). The pattern of duplications within the 14.5 kb subterminal region of the *Oxytricha* mitochondrial genome suggests that successive duplications toward the telomere have resulted in the formation of paralogs and pseudogenes and that these duplications are partly responsible for the larger genome size. The presence of a mitochondrial plasmid that contains a region that matches segmental duplications on the primary mitochondrial genome indicates either an association between this plasmid and the duplications or possibly a higher probability of incorporation of the element in this region of the genome. Such an integration of the plasmid might have only mildly deleterious effects.

Ciliate mitochondrial genomes appear to have high gene densities and are considered to be "large" (Gray et al. 2004) because they contain a relatively large number of mitochondrial genes (29 known protein-coding genes, 11 tRNAs, and 2 rRNAs in *Oxytricha*). The current expanded set of five split mitochondrial genes in ciliates (*nad1*, *nad2*, *rps3*, *lsu*, and *ssu*) suggests that there is no specific functional correlation with the presence of split genes in these genomes. At least some of the gene splits (in *nad1*, *rps3*, and *lsu*) occur at approximately the same gene position and therefore appear to have occurred prior to the last common ancestor of these organisms. Additional ciliate mitochondrial genome sequences may reveal more cases of split genes, which are currently hard to detect due to the extreme divergences of ciliate mitochondrial genomes and relative paucity of sequence data from the broader diversity of ciliates.

It was previously shown that the oligohymenophoran *cox1*, *cox2*, and *cob* genes are extremely divergent relative to other eukaryotic mitochondrial genes, and this is partly responsible for the difficulty in classifying a large number of ciliate mitochondrial ORFs (Burger et al. 2000). Extreme divergence appears to be a general property of ciliate mitochondrial protein-coding genes, even for the highly conserved iron-sulfur proteins *nad7* and *nad10* (the least divergent of the ciliate mitochondrial proteins; supplementary fig. 2, Supplementary Material online). There appears to be no functional association with such extreme divergence because genes with unrelated functions (ribosomal, electron transport, and protein transport/maturation genes) all exhibit extreme divergences. Though ciliate mitochondrial rRNA genes do not appear to be evolving at exceptional rates (Gray and Spencer 1996) in relation to other eukaryotic mitochondrial rRNAs, their distances and divergence rates may be underestimated, due to saturation (supplementary fig. 3, Supplementary Material online). Therefore, gene substitution rates appear to be generally elevated in ciliate mitochondria, irrespective of whether the gene encodes a protein or RNA product. A neutral evolutionary process due to low fidelity replication or error-prone repair would be consistent with the elevated ciliate mitochondrial substitution rates.

There is also evidence to suggest that ciliate nuclear genes have elevated substitution rates relative to that of other eukaryotes, with ciliates such as *Oxytricha* and *Euplotes*—that possess a highly fragmented macronuclear genome structure—evolving the most rapidly (Zufall et al. 2006). Unlike the case of mitochondrially encoded genes, we have not observed evidence of extreme divergences in any of the nuclear-encoded, putatively mitochondrially targeted, genes that we examined in this study (RF1, mtDNA polymerase, mtRNA polymerase [Moriyama et al. 2008], *nad8*, *nad11*). Furthermore, since different DNA polymerase complexes are responsible for nuclear versus mitochondrial replication, we do not expect a correlation between elevated mitochondrial substitution rates and nuclear substitution rates.

Based on comparisons of the oligohymenophorean and spirotrich mitochondrial genomes, we propose that their common ancestor possessed: 1) a linear mitochondrial genome; 2) a replication origin within- or in close proximity to an AT-rich region of low complexity; and 3) TIRs capped by telomeric repeats.

In ciliate mitochondrial genomes, both the putative replication origin (Arnberg et al. 1974; Goddard and Cummings 1975, 1977; Pritchard and Cummings 1981) and primary region of transcription initiation appear to lie in close proximity to, or coincide with, a low-complexity/repeat region. TATA-like elements in multiple *Paramecium* species have been proposed as a motif for transcription recognition (Pritchard et al. 1983) but this now seems unlikely given that a different T-odd phage-like eukaryotic mitochondrial RNA polymerase is most likely the primary mitochondrial RNA polymerase, and such phage RNA polymerases are TATA independent (Cermakian et al. 1996; Shutt and Gray 2006). In *Tetrahymena* species, a highly conserved, GC box-like region in the central region of divergent transcription has been proposed as a motif that may be responsible for initiating transcription and possibly also DNA replication (Moradian et al. 2007). Experimental evidence is necessary to pinpoint the precise location of transcription initiation in these genomes.

Both mitochondrial TIRs and telomeric sequences, such as those of *Tetrahymena* and *Oxytricha*, were proposed to be of foreign origin (Nosek and Tomáška 2003). Nosek and Tomáška (2003) also proposed that linear mitochondrial genomes owe their linearity to mobile elements, which would provide both the need and means to replicate linear genomes by providing DNA sequences/structures and a polymerase necessary for replicating linear DNA. The *Oxytricha* linear mitochondrial plasmid appears to lack the TIRs characteristic of most known linear plasmids (Meinhardt et al. 1990; Handa 2008). Instead, it had a 5′ end with complex repeats and a 3′ end with the same telomeric repeats as the primary mitochondrial genome. The latter feature demonstrates that it is possible to transfer telomeric sequence repeats between mitochondrial genomes and linear plasmids.

One possible scenario for such a transfer is that the original *Oxytricha* linear plasmid may have possessed a terminal inverted structure, which was lost during mitochondrial genome integration, followed by capture of a telomere-bearing end from the primary genome. Alternatively, the plasmid may have possessed a similar structure to its current form, with a telomeric repeat sequence that was transferred to the *Oxytricha* mitochondrial genome during an integration event. We propose that, as for horizontal gene transfer, the phagotrophic lifestyles of ciliates may predispose them to periodic mitochondrial invasions by mobile elements bearing error-prone DNA polymerases, such as the *Oxytricha* mO plasmid. These foreign polymerases may in turn interfere with or partially substitute for the primary, higher fidelity mitochondrial DNA polymerase, contributing to the extreme evolutionary divergences observed in ciliate mitochondria.

Supplementary Material

Supplementary table 1 and figures, 1–3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We would like to thank the Hans Lipps laboratory (Universität Witten/Herdecke), in particular Franziska Jönsson, for providing us with DNA from *Stylonychia* and Jingmei Wang for general laboratory assistance. This research was supported by the National Institutes of Health grant GM59708 to L.F.L.

Literature Cited

Akhmanova A, et al. 1998. A hydrogenosome with a genome. Nature 396(6711):527–528.

Altschul S, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Arnberg AC, et al. 1974. An analysis by electron microscopy of intermediates in the replication of linear *Tetrahymena* mitochondrial DNA. Biochim Biophys Acta. 361(3):266–276.

Barth D, Berendonk TU. 2011. The mitochondrial genome sequence of the ciliate *Paramecium caudatum* reveals a shift in nucleotide composition and codon usage within the genus *Paramecium*. BMC Genomics 12:272.

Binder S, et al. 1992. RNA editing in trans-splicing intron sequences of *nad2* mRNAs in *Oenothera* mitochondria. J Biol Chem. 267(11):7615–7623.

Boxma B, et al. 2005. An anaerobic mitochondrion that produces hydrogen. Nature 434(7029):74–79.

Brunk CF, et al. 2003. Complete sequence of the mitochondrial genome of *Tetrahymena thermophila* and comparative methods for identifying highly divergent genes. Nucleic Acids Res. 31(6):1673–1682.

Burger G, et al. 2000. Complete Sequence of the Mitochondrial Genome of *Tetrahymena pyriformis* and Comparison with *Paramecium aurelia* Mitochondrial DNA. J Mol Biol. 297:365–380.

Burgers PMJ, et al. 2001. Eukaryotic DNA polymerases: proposal for a revised nomenclature. J Biol Chem. 276(47):43487–43490.

Cermakian N, et al. 1996. Sequences homologous to yeast mitochondrial and bacteriophage T3 and T7 RNA polymerases are widespread throughout the eukaryotic lineage. Nucleic Acids Res. 24(4):648–654.

Claros MG, Vincens P. 1996. Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem. 241(3):779–786.

Dawson D, Herrick G. 1982. Micronuclear DNA sequences of *Oxytricha fallax* homologous to the macronuclear inverted terminal repeat. Nucleic Acids Res. 10(9):2911–2924.

de Graaf RM, et al. 2009. The mitochondrial genomes of the ciliates *Euplotes minuta* and *Euplotes crassus*. BMC Genomics 10:514.

de Graaf RM, et al. 2011. The organellar genome and metabolic potential of the hydrogen-producing mitochondrion of *Nyctotherus ovalis*. Mol Biol Evol. 28(8):2379–2391.

Dinouel N, et al. 1993. Linear mitochondrial DNAs of yeasts: closed-loop structure of the termini and possible linear-circular conversion mechanisms. Mol Cell Biol. 13(4):2315–2323.

Drummond AJ, et al. 2009. Geneious v4.7 [Internet]. [cited 2011 September]. Available from: http://www.geneious.com/.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(50):1792–1797.

Eisen JA, et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. 4(9):e286.

Elo A, et al. 2003. Nuclear genes that encode mitochondrial proteins for DNA and RNA metabolism are clustered in the *Arabidopsis* genome. Plant Cell 15(7):1619–1631.

Endoh H, et al. 1994. Hairpin and dimer structures of linear plasmid-like DNAs in mitochondria of *Paramecium caudatum*. Curr Genet. 27(1):90–94.

Fan J, Lee RW. 2002. Mitochondrial genome of the colorless green alga *Polytomella parva*: two linear DNA molecules with homologous inverted repeat Termini. Mol Biol Evol. 19(7):999–1007.

Forget L, et al. 2002. *Hyaloraphidium curvatum*: a linear mitochondrial genome, tRNA editing, and an evolutionary link to lower fungi. Mol Biol Evol. 19(3):310–319.

Goddard JM, Cummings DJ. 1975. Structure and replication of mitochondrial DNA from *Paramecium aurelia*. J Mol Biol. 97(4):593–609.

Goddard JM, Cummings DJ. 1977. Mitochondrial DNA replication in *Paramecium aurelia*. Cross-linking of the initiation end. J Mol Biol 109(2):327–344.

Gray M, et al. 1998. Genome structure and gene content in protist mitochondrial DNAs. Nucleic Acids Res. 26(4):865–878.

Gray M, Spencer D. 1996. Organellar evolution. In: Roberts DMcL, et al. editors. Evolution of microbial life: 54th Symposium of the Society for General Microbiology held at the University of Warwick. Symposia of the Society for General Microbiology; 1996 Mar; Coventry, UK. Cambridge: Cambridge University Press. p. 109–126.

Gray MW, Lang BF, Burger G. 2004. Mitochondria of protists. Ann Rev Genet. 38(1):477–524.

Griffiths-Jones S, et al. 2005. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res. 33(Suppl 1):D121–D124.

Handa H. 2008. Linear plasmids in plant mitochondria: peaceful coexistences or malicious invasions. Mitochondrion 8(1):15–25.

Heinonen T, et al. 1987. Rearranged coding segments, separated by a transfer RNA gene, specify the two parts of a discontinuous large subunit ribosomal RNA in *Tetrahymena pyriformis* mitochondria. J Biol Chem. 262(6):2879–2887.

Hildebrand A, et al. 2009. Fast and accurate automatic structure prediction with HHpred. Proteins Struct Funct Bioinformat 77(Suppl 9):128–132.
Hofacker IL. 2003. Vienna RNA secondary structure server. Nucleic Acids Res. 31(13):3429–3431.
Hopfner K, et al. 1999. Crystal structure of a thermostable type B DNA polymerase from *Thermococcus gorgonarius*. Proc Natl Acad Sci U S A 96(7):3600–3605.
Horton TL, Landweber LF. 2000. Mitochondrial RNAs of myxomycetes terminate with non-encoded 3′ poly(U) tails. Nucleic Acids Res. 28(23):4750–4754.
Iwamoto M, et al. 1998. A ribosomal protein gene cluster is encoded in the mitochondrial DNA of *Dictyostelium discoideum*: uGA termination codons and similarity of gene order to *Acanthamoeba castellanii*. Curr Genet. 33(4):304–310.
Kaguni LS. 2004. DNA Polymerase, the mitochondrial replicase. Ann Rev Biochem. 73(1):293–320.
Kayal E, Lavrov DV. 2008. The mitochondrial genome of *Hydra oligactis* (Cnidaria, Hydrozoa) sheds new light on animal mtDNA evolution and cnidarian phylogeny. Gene 410(1):177–186.
Kazmierczak K, et al. 2002. The phage N4 virion RNA polymerase catalytic domain is related to single-subunit RNA polymerases. EMBO J. 21(21):5815–5823.
Kelley LA, Sternberg MJE. 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc. 4(3):363–371.
Kempken F, Hermanns J, Osiewacz H. 1992. Evolution of linear plasmids. J Mol Evol. 35(6):502–513.
Krogh A, et al. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305(3):567–580.
Lecrenier N, Foury F. 2000. New features of mitochondrial DNA replication system in yeast and man. Gene 246(1–2):37–48.
Lecrenier N, van der Bruggen P, Foury F. 1997. Mitochondrial DNA polymerases from yeast to man: a new family of polymerases. Gene 185(1):147–152.
Lowe T, Eddy S. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.
Malek O, Brennicke A, Knoop V. 1997. Evolution of trans-splicing plant mitochondrial introns in pre-Permian times. Proc Natl Acad Sci U S A. 94(2):553–558.
Markham NR, Zuker M. 2005. DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res. 33(Suppl 2):W577–W581.
Marriott AC, Archer SA, Buck KW. 1984. Mitochondrial DNA in *Fusarium oxysporum* is a 46.5 kilobase pair circular molecule. J Gen Microbiol. 130(11):3001–3008.
Massey SE, Garey JR. 2007. A comparative genomics analysis of codon reassignments reveals a link with mitochondrial proteome size and a mechanism of genetic code change via suppressor tRNAs. J Mol Evol. 64(4):399–410.
Meinhardt F, et al. 1990. Linear plasmids among eukaryotes: fundamentals and application. Curr Genet. 17(2):89–95.
Meinhardt F, Schaffrath R, Larsen M. 1997. Microbial linear plasmids. Appl Microbiol Biotechnol. 47(4):329–336.
Moradian MM, et al. 2007. Complete mitochondrial genome sequence of three *Tetrahymena* species reveals mutation hot spots and accelerated nonsynonymous substitutions in ymf genes. PLoS One 2(7):e650.
Mori Y, et al. 2005. Plastid DNA polymerases from higher plants, *Arabidopsis thaliana*. Biochem Biophys Res Commun. 334(1):43–50.
Morin GB, Cech TR. 1986. The telomeres of the linear mitochondrial DNA of *Tetrahymena thermophila* consist of 53 bp tandem repeats. Cell 46(6):873–883.
Morin GB, Cech TR. 1988. Mitochondrial telomeres: surprising diversity of repeated telomeric DNA sequences among six species of *Tetrahymena*. Cell 52(3):367–374.
Moriyama T, et al. 2008. Purification and characterization of organellar DNA polymerases in the red alga *Cyanidioschyzon merola*. FEBS J. 275(11):2899–2918.
Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25(10):1335–1337.
Nosek J, Tomáška Ľ. 2003. Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy. Curr Genet. 44(2):73–84.
Nosek J, et al. 1998. Linear mitochondrial genomes: 30 years down the line. Trends Genet. 14(5):184–188.
Nowacki M, et al. 2009. A functional role for transposases in a large eukaryotic genome. Science 324(5929):935–938.
Pritchard AE, Cummings DJ. 1981. Replication of linear mitochondrial DNA from *Paramecium*: sequence and structure of the initiation-end crosslink. Proc Natl Acad Sci U S A. 78(12):7341–7345.
Pritchard AE, et al. 1983. Inter-species sequence diversity in the replication initiation region of *Paramecium* mitochondrial DNA. J Mol Biol. 164(1):1–15.
Pritchard AE, et al. 1990. Nucleotide sequence of the mitochondrial genome of *Paramecium*. Nucleic Acids Res. 18(1):173–180.
Ricard G, et al. 2008. Macronuclear genome structure of the ciliate *Nyctotherus ovalis*: single-gene chromosomes and tiny introns. BMC Genomics 9(1):587.
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16(6):276–277.
Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng. 12(2):85–94.
Sakurai R, et al. 2000. *In vivo* conformation of mitochondrial DNA revealed by pulsed-field gel electrophoresis in the true slime mold, *Physarum polycephalum*. DNA Res. 7:83–91.
Schnare MN, Greenwood SJ, Gray MW. 1995. Primary sequence and post-transcriptional modification pattern of an unusual mitochondrial tRNAMet from *Tetrahymena pyriformis*. FEBS Lett. 362(1):24–28.
Schnare MN, et al. 1986. A discontinuous small subunit ribosomal RNA in *Tetrahymena pyriformis* mitochondria. J Biol Chem. 261(11):5187–5193.
Scolnick E, et al. 1968. Release factors differing in specificity for terminator codons. Proc Natl Acad Sci U.S.A. 61(2):768–774.
Seilhamer JJ, Gutell RR, Cummings DJ. 1984. *Paramecium* mitochondrial genes. II. Large subunit rRNA gene sequence and microevolution. J Biol Chem. 259(8):5173–5181.
Seilhamer JJ, Olsen GJ, Cummings DJ. 1984. *Paramecium* mitochondrial genes. I. Small subunit rRNA gene sequence and microevolution. J Biol Chem. 259(8):5167–5172.
Sethuraman J, et al. 2009. Molecular evolution of the mtDNA encoded rps3 gene among filamentous ascomycetes fungi with an emphasis on the Ophiostomatoid fungi. J Mol Evol. 69(4):372–385.
Shutt TE, Gray MW. 2006. Bacteriophage origins of mitochondrial replication and transcription proteins. Trends Genet. 22(2):90–95.
Small I, et al. 2004. Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4:1581–1590.

Smith DG, et al. 2007. Exploring the mitochondrial proteome of the ciliate protozoon *Tetrahymena thermophila*: direct analysis by tandem mass spectrometry. J Mol Biol. 374(3):837–863.

Smits P, et al. 2007. Reconstructing the evolution of the mitochondrial ribosomal proteome. Nucleic Acids Res. 35(14):4686–4703.

Soding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7):951–960.

Soding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 33(Suppl 2):W244–W248.

Suyama Y, Miura K. 1968. Size and structural variations of mitochondrial DNA. Proc Natl Acad Sci U S A. 60:235–242.

Takano H, Kawano S, Kuroiwa T. 1992. Constitutive homologous recombination between mitochondrial DNA and a linear mitochondrial plasmid in *Physarum polycephalum*. Curr Genet. 22(3):221–227.

Takano H, Kawano S, Kuroiwa T. 1994. Complex terminal structure of a linear mitochondrial plasmid from *Physarum polycephalum*: three terminal inverted repeats and an ORF encoding DNA polymerase. Curr Genet. 25(3):252–257.

Tsukii Y, Endoh H, Yazaki K. 1994. Distribution and genetic variabilities of mitochondrial plasmid-like DNAs in *Paramecium*. Jpn J Genet. 69(6):685–696.

UniProt Consortium. 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 39(Suppl 1):D214–D219.

Vahrenholz C, et al. 1993. Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. Curr Genet. 24(3):241–247.

Walther TC, Kennell JC. 1999. Linear mitochondrial plasmids of *F. oxysporum* are novel, telomere-like retroelements. Mol Cell. 4(2):229–238.

Wan F, et al. 2007. Ribosomal protein S3: a KH domain subunit in NF-[kappa]B complexes that mediates selective gene regulation. Cell 131(5):927–939.

Will S, et al. 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol. 3(4):e65.

Wright ADG, Lynn DH. 1997. Maximum ages of ciliate lineages estimated using a small subunit rRNA molecular clock: crown eukaryotes date back to the Paleoproterozoic. Arch Protistenkd. 148(4):329–342.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13(5):555–556.

Zufall RA, et al. 2006. Genome architecture drives protein evolution in ciliates. Mol Biol Evol. 23(9):1681–1687.

Xia X, Xie Z. 2001. DAMBE: data analysis in molecular biology and evolution. J Hered. 92:371–373.

Associate editor: Shu-Miaw Chaw

Genome Biology and Evolution

SMBE